1. caGRID System Architecture Recap

▶ Requirement analysis:
  – Promote security to be visible at overview level.
  – Detail roles, specially for CTMS.
  – Versioning: Detail requirements. Define place holder.
  – Early for feedback, need to review documentation. Feedback from Arch/V-CDE workspace members.
  – CTMS and TBPT to Detail de-identification and security.

▶ Architecture:
  – Work with ID, Workflow SIGs.
  – Aware of technology evolution.

▶ Implementation:
  – Expectation management about what will be implemented.

▶ Process/methodology:
  – Audience agrees with the proposed process – Use case analysis/technology evaluation/system architecture/Reference implementation.

2. Workflow Recap

▶ Presentation summary
- ICR use cases
- Technology overview
- Grid architecture
- Issues facing workflow

▶ Conclusions
- Workflow is indeed important in many ways
- We should consider some different use cases
  - Clinical trials workflow
  - Workflow composition (esp. metadata/provenance issues)
- We should consider technologies from different domains
  - Process and resource scheduling
- We should move forward with reference implementations and white papers
  - Workflow can be developed in parallel with caGRID
  - Evaluate all technologies
    - Don't forget WFMC
    - Evaluate technologies with real implementations
  - Get use cases from domain workspaces
  - Provide white paper

3. Metadata Representation: Take Home Points

▸ Metadata should be a separate service provided by any node on the Grid.

▸ Metadata will be represented in a multi-tier fashion.
  – The first tier will be an XML schema describing the structure of the data object.
  – The next tier will contain the data object's metadata from the caDSR. The proposed representation is an XML schema. We propose conducting a review in the next 2-3 weeks to determine if XML schema is sufficient.
  – The final tier will contain the semantic information represented in OWL/OWL-S.

▸ A minimal set of metadata needs to be identified for data and analytical service providers. The caGRID phase I white paper suggests a starting point that can be used to devise these lists.

4. Data Provenance: Take Home Points

▸ caBIG should adopt a hierarchical provenance model

▸ Provenance information should be attached to any persistent data store; however, the data store is only required to provide provenance information to its immediate source of information, not all the way back to the original source.

▸ The provider determines the type of information that constitutes its provenance. This should be sufficient information for an interested party to find the source information and understand the transformation that was performed on the data.

▸ Analytical services should provide provenance information as the data passes through workflows.

▸ A single group should be formed to work on ID, versioning and provenance.

5. ID Management Framework Recap

▸ Attendees at W3C Workshop will summarize results and communicate to group (via Arumani)

▸ Bob Robbins and Harold Solbrig (and others) will convene via e-mail and teleconferences over the next two weeks to develop coherent assessment and recommendations for group

▸ Individuals interested in participating over the next two weeks should notify Arumani

▸ Those interested in versioning and provenance issues should coordinate with the identifier group (join)

6. Common Query Language Recap

▶ A single common language for clients, instead of multiple specialized languages.

▶ A language based on object-oriented data source views

▶ Need to look at how ability to query over semantic/ontology information will be integrated into the language
  – A separate set of constructs and language extensions may be needed

▶ Separation of query interface from query execution engine
  – A well defined set of interfaces for clients to use.
  – A well defined set of interfaces for data source access
  – Middleware service to map queries defined in the common language to an execution plan.

▶ Need to look at how to access analytical services

▶ Joins will be important
  – Use of foreign keys, identifiers to support efficient execution of equality joins
  – Do not want to limit ourselves only to equi-joins

▶ Technologies to start from
  – caBIO
  – XQuery/XPath
  – OQL
  – DQP
  – RDQL (for semantic information and ontologies)

Common Query Language Recap (cont'd)

▶ A phased implementation
- – Start from support for object level queries
- – Support for equi-joins across disparate data sources
- – Incorporate semantic information related to objects
- – Incorporate semantic information about relationships
- – Support for service composition.

7. HL7 Messaging Recap

▸ HL7 as a standard will need to be addressed as it applies to caBIG from several perspectives(Federal requirements, support of current HL7 users, reuse of previously produced specifications that apply)

▸ HL7 does not satisfy all data messaging requirements for caBIG, but may apply to clinical data messaging

▸ HL7 v3 messaging use cases need to be defined by Clinical Trials workspace and ICR

▸ HL7 data types may be applicable outside of the clinical domain

▸ HL7 V3 is not in wide use and V2 will need to be supported, but Workspace developers should be encouraged to consider V3 because of semantic richness

▸ Final decisions on applicability of HL7 requires further study

8. Data and Vocabulary Standards Recap

▸ General agreement to caBIG data standards governance process with minor modifications based on feedback.

▸ V-CDE to develop detailed process for review, approval, and maintenance of data standards.

▸ V-CDE will consider procedures for notification of caBIG members regarding changes to data standards.

▸ Action Item:

– Kathleen Gundry will provide V-CDE workspace with advanced copies of data standards under consideration by the NCI Context Administrators as possible template for data standards submission package.

9. Vocabulary Deployment Recap

▸ Centralized vs. Distributed Terminology Servers
  – Do we support caching of terminologies?

▸ Do we allow local extensions?

▸ Do we have an infrastructure to support development of local vocabularies?